A Method for Extracting Causal Knowledge from Textual Databases

CHRISTOPHER KHOO, SYIN CHAN, YUN NIU & ALYSSA ANG

Abstract

This paper describes the first phase of a project to develop a knowledge extraction and knowledge discovery system that extracts causal knowledge from a textual database automatically, and attempts to infer new causal relationships from the extracted information. The initial work is focused on developing an automatic method for identifying and extracting cause-effect information expressed in medical abstracts. Linguistic clues that indicate the presence of a causal relation in text are being identified, and linguistic patterns constructed to represent the different ways in which cause and effect are expressed in English sentences. The linguistic patterns have "slots" that indicate the parts of the sentence representing the cause and the effect. The information extraction process involves matching the linguistic patterns with the syntactic structure of sentences, and extracting the parts of the sentence that match with the slots in the patterns. The extracted information is stored in a structured manner in a "cause-effect template" that indicates the different roles and attributes of the causal situation described in the text.

Introduction

This paper describes the first phase of a project to develop a knowledge extraction and knowledge discovery system that extracts causal knowledge from textual databases automatically, and attempts to infer new causal relationships from the extracted information. Our initial work is focused on developing an automatic method to extract cause-effect information automatically from the MEDLINE database that contains abstracts of medical articles. Extraction of cause-effect information from text is a particular type of information extraction. Information extraction deals with the problem of identifying and extracting the text fragments in a document that answer a particular question of interest or

that describe aspects of a particular type of event or concept (Message Understanding Conference, 1992, 1994, 1995 & 1998; Cardie, 1997; Cowie & Lehnert, 1996).

The information extracted from the text is usually represented in a structured manner, in the form of a template with a set of slots. Each slot indicates an attribute or aspect of an event or concept, and the whole template when filled describes the event or concept in a concise way. The process of information extraction involves filling in the slots with information expressed in the text. The information thus extracted can be used for building a specialized factual database that can be searched easily, for data mining, for use in a knowledge-base of an expert system, or for use in automatic summarization.

Previous studies in information extraction have focused on extracting information relevant to a particular type of event, e.g. terrorist attacks and business joint ventures (Message Understanding Conference, 1992 & 1994), or to a particular application area, e.g. summarizing patient medical records (Soderland et al., 1995) and analyzing life insurance applications (Glasgow et al., 1998). The focus of this study is on extracting a particular type of relation – the cause-effect relation – and a particular type of knowledge –the knowledge of causal relationships.

However, our approach is similar to those used in other information extraction studies. We first construct a cause-effect template to represent the different roles and attributes in a causal situation. We then identify the variety of ways in which the causal relation is expressed in text. We represent the different ways of expressing cause-effect as linguistic patterns with slots to be filled in. Any part of the text that matches a particular pattern is considered to contain a causal relation, and the words in the text that match the slots in the pattern are extracted and used to fill the appropriate slots in the cause-effect template.

We foresee that information extraction will be an important application in digital libraries. It can be used for the following purposes:

1. To improve information retrieval effectiveness. Traditional approaches of information retrieval identify documents that contain the keywords specified in a query. Information extraction techniques can be used to identify documents that contain information sought by the user or information that answers the user's query.
2. To construct a conceptual map of the library for information visualization. Information extracted from a digital library can be chained or connected to give an overview of the information available in the digital library. The conceptual map can show how information in one document is related to information in another document, and how the documents are related in their contents.

3. To synthesize new knowledge. New knowledge can be synthesized by connecting or chaining the pieces of information extracted from related documents.
4. To support creativity by suggesting hypotheses for investigation and indicating gaps in knowledge. When two pieces of information extracted from the digital library are combined, the potential new knowledge can be treated as a hypothesis to be investigated. It may also be possible to analyze the conceptual map and identify knowledge gaps and promising areas for investigation.

We believe that the extraction of cause-effect information is particularly important for the above purposes. After all, most research studies have the ultimate goal of discovering causal relations between different factors and events.

The concept of causality is surprisingly difficult to define (Khoo, Chan & Niu, in press). Philosophers have grappled with the concept for millennia. A traditional definition can be found in the *Modern Dictionary of Sociology* (Theodorson & Theodorson, 1979) as follows:

- An event (or events) that precedes and results in the occurrence of another event. Whenever the first event (the cause) occurs, the second event (the effect) necessarily or inevitably follows. Moreover, in simple causation the second event does not occur unless the first event has occurred. Thus the cause is both the sufficient condition and the necessary condition for the occurrence of the effect.
- With the conception of multiple causation, various possible causes may be seen for a given event, any one of which may be a sufficient but not necessary condition for the occurrence of the effect, or a necessary but not sufficient condition. In the case of multiple causation, then, the given effect may occur in the absence of all but one of the possible sufficient but not necessary causes; and, conversely, the given effect would not follow the occurrence of some but not all of the various necessary but not sufficient causes.

This paper describes the cause-effect template that we have developed to represent and store the cause-effect information extracted from text, as well as the approach we are taking to extract the information automatically. The text collection used in this study consists of abstracts obtained from the MEDLINE and Psychological Abstracts databases. The emphasis is on extracting cause-effect information that is explicitly expressed in the text, with the minimum use of knowledge-based inferences. It is hoped that this will result in a method that is more easily portable to other subject areas and document collections.

Previous Studies

Most studies on the automatic extraction of causal knowledge from text make use of knowledge-based inferences to infer the causal relations. These studies have focused on the following kinds of text:

1. episodic or narrative text, which describes a series of related events involving human actions (e.g. a story);
2. short explanatory messages that a human might enter into a computer system as part of a human-computer dialog on a particular subject;
3. expository text of the kind found in textbooks.

Research with episodic text seeks to develop computer programs that perform comprehension tasks like answering questions about stories and summarizing the stories (e.g. Bozsahin & Findler, 1992; Mooney, 1990; Schank, 1982; Schubert & Hwang, 1989). These studies attempt to discover the kinds of knowledge and inferences that are needed to identify causal relations between events described in the text and to infer events that are implied in the text. These studies typically make little use of linguistic clues to identify causal relations.

Selfridge, Daniell and Simmons (1985) and Joskowsicz, Ksiezyk and Grishman (1989) have developed prototype computer programs that extract causal knowledge from short explanatory messages entered into the knowledge acquisition component of an expert system. When there is an ambiguity concerning whether a causal relation between two events is expressed in the text, the system uses the domain model to check whether such a causal relation between the events is possible.

Kontos and Sidiropoulou (1991) and Kaplan and Berry-Rogghe (1991) have worked with scientific texts. They used linguistic patterns to identify causal relations, but all the information required for linguistic processing – the grammar, the lexicon, and the patterns for identifying causal relations – were hand-coded and were developed just to handle the sample texts used in the studies. Knowledge-based inferences were also used. The authors pointed out that a substantial amount of subject knowledge, which had to be specified manually, was needed for the system to identify causal relations in the sample texts accurately. Scaling up is obviously a problem: the grammar, lexicon and patterns will not be usable in another subject area, and may not even be effective for other documents on the same subject.

More recently, Garcia (1997) developed a computer program to extract cause-effect information from French technical texts without using domain knowledge. He focused on causative verbs and reported a precision rate of 85%. Khoo, Kornfilt, Myaeng & Oddy (1998) developed an automatic method for extracting cause-effect information from Wall Street Journal texts using linguis-

tic clues and pattern matching. Their method has successfully extracted about 68% of the causal relations with an error rate of about 36%. The modest results were obtained probably because *Wall Street Journal* articles are non-technical and cover a very wide range of topics.

Our current study, which also makes use of linguistic clues of causality, and pattern matching, focuses on abstracts from MEDLINE and Psychological Abstracts. It is hoped that more accurate and useful results will be obtained with these databases because the causal relation is important in the field of medicine and experimental psychology, and is more likely to be explicitly expressed in the abstracts.

A major motivation for this study is the ability to synthesize new knowledge from the causal knowledge extracted from the document collection. In a series of studies, Swanson (1986) has demonstrated that logical connections between the published literature of two research areas related to medicine can provide new and useful hypotheses. Suppose an article reports that A causes B, and another article reports that B causes C, then there is an implicit logical link between A and C (i.e. A causes C). This relation would not become explicit unless work is done to extract it. This gives rise to the idea that new discoveries can be made by analyzing published literature (Finn, 1998).

Swanson has proposed uncovering these implicit connections using information retrieval techniques (Swanson & Smalheiser, 1997). To facilitate the process of identifying the "influence" and "similarity" relationship between A-B and B-C, Swanson and Smellaiser (1998) developed a computer program called ARROWSMITH (http://kiwi.uchicago.edu/). However, the program has several limitations:

1. It operates only on the article's title and not on the full text or abstract.
2. It requires a significant investment in time and effort from its users. Users must first conduct two detailed MEDLINE searches, download the results in a specific text format, and then upload the results to the ARROWSMITH site for processing.

Our study seeks to extract causal knowledge from text using natural language processing and information extraction techniques so that the knowledge discovery process can be more automated.

The Cause-Effect Template

The cause-effect template for representing causal knowledge was developed from an analysis of about 100 abstracts, and given in Table 1. The various kinds of information that are relevant to a causal situation are identified and

represented in the cause-effect template. The text might not contain all the information listed in the template, and so not all the slots need be filled during information extraction. An asterisk (*) indicates that the role can occur a multiple number of times. For example, there may be a conjunction of several causes for an effect, or there may be multiple effects. The major roles in the cause-effect template are *cause, effect, condition, modality, evidence, linguistic expression* and *type of causal relation*.

Condition specifies the environment for the cause to produce the effect. The difference between *condition* and *cause* is fuzzy because an effect is usually the result of several causal factors. So, one causal factor can be said to provide a favourable environment for other factors to produce an effect. However, one of the causal factors is usually highlighted or given prominence as *the cause*, and the other factors are referred to as *conditions*.

Modality indicates the extent to which the causal relation is true or false. *Evidence* is the evidence supplied in the text for supporting the causal relation. *Linguistic expression* is strictly speaking not a role in the causal relation but rather it is the linguistic means used by the author of the text to express the causal relation. *Type of causal relation* indicates whether the *cause* is a mechanical / physical cause, mental / psychological cause, teleological or final cause, or some other type of causality. Some of these major roles can be decomposed into sub-roles. These are also listed in Table 1.

The roles *cause, effect* and *condition* have the sub-roles *object, state/event* and *size*. *Object* refers to a person or thing that causes something or that experiences the effect. The object can be abstract or concrete, and can be vague, general or explicit. *State/event* refers to the relevant aspect of the object that produces the effect or that is changed by the cause. *State* is the condition in which the object is in, and this includes attributes of the object. *Event* is a happening involving the object. The following are some sample sentences specifying these roles:

> The [patient]$^{effect.object}$ is [not feeling well]$^{effect.state}$ because of [stomachache]^{cause} due to overeating.
> The [patient]$^{effect.object}$ is not feeling well because of [stomachache]$^{effect.state}$ due to [overeating]$^{cause.event}$.

Table 1. The Cause-Effect Template

Level 1	Level 2	Level 3
Cause: *	Object:	Value: Type:
	State / Event:	Value: Type:
	Size:	Sufficient condition: Minimum condition: Maximum condition:
Effect: *	Object:	Value: Type:
	State / Event:	Value: Type:
	Size:	Strength: Percentage / Number: Comparison.Before: Comparison.After: Compari- son.Greater_than: Comparison.Less_than: Comparison.Same_as: Equation:
	Polarity: "In- crease | Decrease | Im- prove | Worsen | Regulate | Prevent | Faster | Slower"	
Condition: *	Object:	Value: Type:
	State / Event:	Value: Type:
	Size:	Sufficient condition: Minimum condition: Maximum condition:
	Duration:	Value: Type:
	Degree of necessity:	
Modality:	Truth value: "True | False | Probable | Possible | Unlikely"	
	Linguistic expression:	
Evidence:	Research method:	
	Sample size:	
	Significance level:	
	Source of information:	
	Location:	
Linguistic expression:		
Type of causal relation:		

* indicates that the slot can occur multiple times

> [Doxycycline sclerotherapy][cause.object] can be used [effectively][effect.polarity:improve] for [pleurodesis][effect.event] in the management of [nontraumatic pneumothorax][condition.state] in the [patient][effect.object] with [AIDS][condition.state].

> In [guinea pigs][condition.object] infected with [HSV][condition.state], subsequent [administration][cause.event] of [ALA][cause.object] and [exposure][cause.event] of the [lesions][cause.object] to [red light][cause.event] [shortened][effect.polarity:decrease] the [duration][effect.event] of [vesicles'][effect.object] [appearance][effect.event] from [more than a week][effect.size.comparison.before] to [a few days][effect.size.comparison.after] and [reduced][effect.polarity:decrease] [HSV titer in the lesions][effect.object] by [> or = 5 log10][effect.size.strength].

Sometimes, the text may not indicate the specific object, state, or event but only the type of object, state or event, e.g. "economic effect", "social impact", "political condition". Hence, object and state/event slots are subdivided into two slots: *value* for specifying the actual object, state or event, and *type* for specifying the type of object, state or event. For example:

> Our desire to understand the [potential][modality:possible] [adverse][effect.polarity:worsen] [human][effect.object] [health effects][effect.state.type] of [environmental chemical exposure][cause.event.type] …

> There is a public perception that [chemicals][cause.object.type] [generally][modality] cause [immunosuppression][effect.state].

Size is an indication of the magnitude of the cause, effect or condition. This can be expressed quantitatively (in which case, the unit of measurement should be specified) or qualitatively using words such as "big", "small", etc. For the *cause* and *condition* roles, *size* is subdivided into *sufficient condition*, *minimum condition* and *maximum condition*. The text might specify the sufficient condition for the effect to occur but that might not be the smallest condition for the effect to occur. There may also be a maximum size, beyond which an "overdose" condition might result. The *effect size* can be expressed as a quantity of the effect or as a percentage of the population (or number of people) experiencing the effect. Some examples of sentences that specify the size of the cause, condition and effect are:

> [Two][effect.size.number] [patients][effect.object] with the [acquired immune deficiency syndrome][condition.state] developed [acute pulmonary edema][effect.event] following [intravenous fluid administration][cause.event].

> It has been [estimated][modality] that [35 million][effect size.number] [Americans][effect.object] [suffer from allergic disease][effect.state], of which [2-5%][effect size.percentage] are from [occupational exposure][cause.event].

The size of an effect may also be expressed as a mathematical relationship between the size of the cause and the size of the effect. *Size* may also be ex-

pressed comparatively, e.g. the effect may be bigger for one drug or one dosage than another. The *condition* role has a *degree of necessity* attribute indicating to what extent it is a necessary condition for the cause to have the stated effect.

Polarity indicates the direction of the effect. *Polarity* can be expressed as an *increase* or *decrease* in something, or an *improvement* or *deterioration* of a situation. For example:

> [Cholinergic agonists][cause.object] have been reported to [ameliorate][effect.polarity:decrease] [ECT-induced memory impairment][effect.state].
>
> [This accumulation][effect.event] was [enhanced][effect.polarity:increase] [approximately two-fold][effect size.strength] in the presence of an [iron chelator][cause.object].
>
> The [use][cause.event] of [polypharmacotherapy][cause.object] in the [treatment][effect.polarity:improve] of [psychiatric disorder][effect.state] is [commonplace][evidence].

Modality or degree of confidence refers to the truth value of the relation. Possible values include *true, false, possible, probable* and *unlikely*. For example:

> [Seven][effect.size.number] [patients][effect.object] [(28%)][effect size.percentage] reported [experiencing symptoms][effect.event] that [could have been][modality.truth_value:possible] caused by one or more of the [herbal products][cause.object.type] that they were taking.

Evidence gives indication of how trustworthy the information is. The source of information may be provided in the text. The text may also provide some supporting statistics such as sample size and significance level, or provide other supporting evidence. For example,

> [Inositol][cause.object] [6 g daily][cause.size] was given in a [cross-over double-blind manner][evidence.research_method] [for 5 days][cause.size] [before the 5th or 6th ECT][condition.event] in a series of [patients][effect.object]. [No effect][effect.size.strength] was found on [post-ECT cognitive impairment][effect.state].
>
> According to [FDA regulations][evidence.source_of_information], a [combination drug][cause.object] is not efficacious unless [each component][condition.object] [contributes to the claimed effects][condition.event].

Different instantiations of the cause-effect template may be related. For example, cause and effect can be expressed at different levels of generality. One instance of a causal relation can be a special case (i.e. a more specific case) of another causal relation. Different instantiations of the cause-effect template may share the same cause or condition. The effect in one template may be the cause in another template. There may also be a comparison of the effect size for different causes (e.g. efficacy of different drugs for treating the same disease). So there can be a web of related cause-effect templates.

We are further developing the template in the following ways:

- to identify additional roles and slots in the cause-effect template.
- to distinguish the optional slots (that need not be filled in) from the mandatory slots.
- to determine the default value for some slots. For example, the *modality.truth_value* slot may have the default value of *true*.
- to determine the complete list of possible values for some slots (e.g. modality).
- to determine the relationships between the slots. For example, the default value for *condition.object* is the same as the value for *cause.object*.

Automatic Extraction of Causal Knowledge

Our approach for extracting causal knowledge is to first parse each sentence using Conexor's Functional Dependency Grammar of English parser (FDG parser) (http://www.conexor.fi) to generate a graph representing the syntactic structure of the sentence. We are developing a set of graphical patterns that specifies the various ways a causal relation can be expressed in text. Each graphical pattern represents one way that the causal relation can be expressed, and it contains slots for the different roles in the cause-effect template.

The information extraction process involves matching the graphical patterns with the graphical representation of the sentence structure. If there is a complete match, then a causal relation is considered to be found. The parts of the sentence structure that match the slots in the pattern are extracted and placed in the appropriate slots of a cause-effect template.

Take as an example the following sentence:

> *A removable prosthesis and a fixed partial denture are used to improve a little girl's appearance and oral function.*

The syntactic structure of the sentence as output by the FDG parser is given in Fig. 1 in a graphical diagram. The syntactic structure can also be represented in

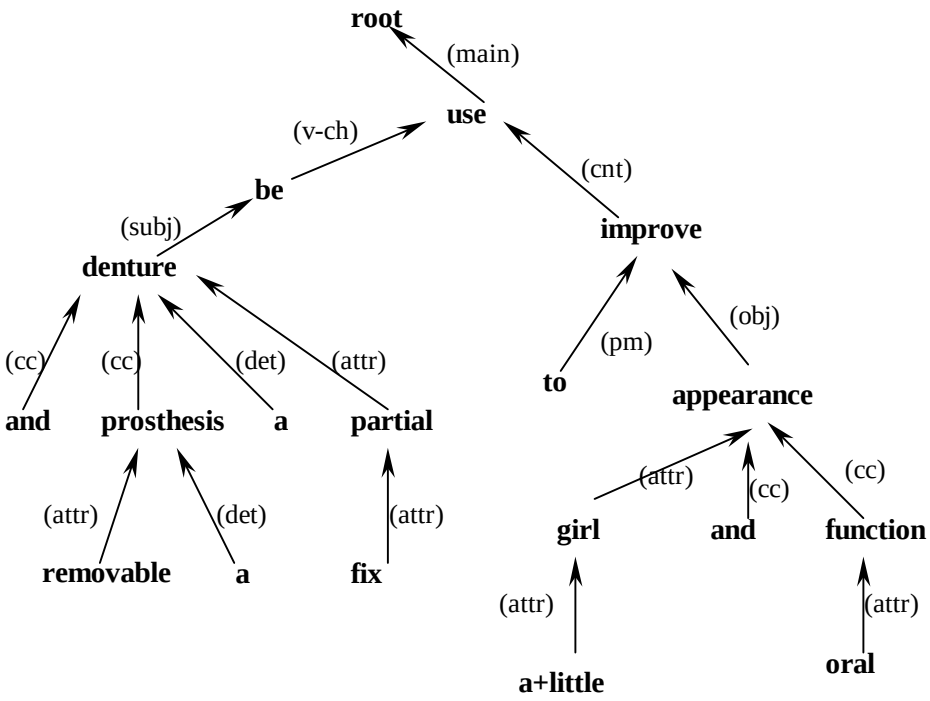

Fig. 1. Syntactic structure of a sentence

linear *conceptual graph* notation (Sowa, 1984) as follows:

```
[use]-
    (v-ch)->[be]->(subj)->[denture]-
                (cc)->[prosthesis]-
                        (det)->[a]
                        (attr)->[removable]  ,
                (cc)->[and]
                (det)->[a]
                (attr)->[partial]->(attr)->[fix]  ,
    (cnt)->[improve]-
            (pm)->[to]
            (obj)->[appearance]-
                    (attr)->[girl]->(attr)->[a+little]
                    (cc)->[and]
                    (cc)->[function]->(attr)->[oral] , , , .
```

In this notation, the concept nodes, representing words in the sentence, are in square brackets and the relation nodes, representing syntactic relations, are

given in parentheses. The arrows (indicating the direction of the relations) point in the reverse direction to the arrows in Fig 1 because of the convention adopted in the conceptual graph notation.

The following linguistic pattern can be used to extract the causal information from the above sentence:

(1) [* : *cause*] <-(subj)<- [“be”] <-(v-ch)<- [“use”] ->(cnt)-> [improve : *effect.polarity=“improve”*] ->(obj)-> [* : *effect*]

Words in quotation marks are stemmed words that must occur in the sentence. Words not in quotation marks (e.g. the word *improve* in the example pattern) refers to a class of synonymous words that can occupy that node. “*” is a wildcard character that can match any term. The roles or slots in the cause-effect template are indicated within the square brackets after the “:” symbol.

The parts of the example sentence that match with the linguistic pattern is indicated in bold below:

[use]-
(v-ch)->[be]->(subj)->[denture]-
(cc)->[prosthesis]-
(det)->[a]
(attr)->[removable] ,
(cc)->[and]
(det)->[a]
(attr)->[partial]->(attr)->[fix] ,
(cnt)->[improve]-
(pm)->[to]
(obj)->[appearance]-
(attr)->[girl]->(attr)->[a+little]
(cc)->[and]
(cc)->[function]->(attr)->[oral] , , , .

“Denture” is extracted as the *cause*, “appearance” is extracted as the *effect*, and “improve” is taken to be the *effect.polarity*.

We have found the need to extend the patterns in two ways:

1. to allow a wildcard to match a subtree (rather than just a single word) in the syntactic structure of the sentence
2. to specify that a node in the pattern can be replaced by a set of equivalent subgraphs.

For example, pattern (1) above can be reformulated as follows:

(2) [&SUBTREE : *cause*] <-(subj)<- [&SUBPATTERN1] ->(obj)->
[&SUBTREE : *effect*]

The concept node [&SUBTREE : *cause*] will match not just a single word (e.g. "denture") but any subtree (e.g. "a removable prosthesis and a fixed partial denture").

The subpattern "(subj)<- [SUBPATTERN1] ->(obj)" can be defined to be equivalent to the following set of subpatterns:

(3) (subj)<- [improve : *effect.polarity*="*improve*"] ->(obj)
(4) (subj)<- ["be"] <-(v-ch)<- ["use"] ->(cnt)->
[improve : *effect.polarity*="*improve*"] ->(obj)
(5) (subj)<- ["will"] <-(v-ch)<- [improve : *effect.polarity*="*improve*"] ->(obj)

So, whenever *&SUBPATTERN1* occurs in a pattern, it can be replaced with subpatterns (3), (4) and (5). This allows pattern (2) to match the following sentences:

A removable prosthesis improves appearance.
A removable prosthesis is used to improve appearance.
A removable prosthesis will improve appearance.

Construction of linguistic patterns for extracting cause-effect from medical abstracts is in progress. We have constructed 24 patterns covering the most common ways of expressing cause and effect. These patterns account for about 70% of the instances of cause-effect in a sample of 200 abstracts. When the 24 patterns were applied to a new sample of 100 abstracts, about 40% of the cause-effect information was correctly extracted. We found it encouraging that a small number of patterns can extract a sizable proportion of cause-effect information. Our aim is to construct a set of patterns to handle at least 80% of the instances of cause and effect.

Conclusion

We have described a method for performing automatic extraction of causal knowledge from textual documents. We use a parser to identify the syntactic structure of a sentence. The structure is matched with a set of graphical patterns that express causal relations. When a match is found, various attributes of the causal situation (e.g. the cause, the effect, the subjects involved and the degree of the effect, etc.) can then be extracted. We have constructed linguistic patterns for the most common ways of expressing cause and effect in medical abstracts.

We have also formulated a cause-effect template that expresses the extracted causal knowledge in a structured manner. We hope that the causal information extracted from medical abstracts and stored in these templates can eventually be used to synthesize new medical knowledge.

References

Bozsahin, H. C., & Findler, N. V. (1992). Memory-based hypothesis formation: Heuristic learning of commonsense causal relations from text. *Cognitive Science, 16*(4), 431-454.

Cardie, C. (1997). Empirical methods in information extraction. *AI Magazine, 18*(4), 65-79.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80-91.

Finn, R. (1998). Program uncovers hidden connections in the literature. *The Scientist, 12*(10), 12-13.

Garcia, D. (1997). COATIS, an NLP system to locate expressions of actions connected by causality links. In E. Plaza & R. Benjamins (Eds.), *Knowledge Acquisition, Modeling and Management: 10th European Workshop, EKAW '97 Proceedings* (pp.347-352). Berlin: Springer-Verlag.

Glasgow, B., Mandell, A., Binney, D., Ghemri, L., & Fisher, D. (1998). MITA: an information-extraction approach to the analysis of free-form text in life insurance applications. *AI Magazine, 19*(1), 59-71

Joskowsicz, L., Ksiezyk, T., & Grishman, R. (1989). Deep domain models for discourse analysis. In H.J. Antonisse, J.W. Benolt, & B.G. Silverman (Eds.), *The Annual AI Systems in Government Conference* (pp. 195-200). Silver Spring, MD: IEEE Computer Society.

Kaplan, R. M., & Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition, 3*(3), 317-37.

Khoo, C., Chan, S., & Niu, Y. (In press). The many facets of the cause-effect relation. In *Semantics of relations.* (To be published by Kluwer.)

Khoo, C.S.G., Kornfilt, J., Oddy, R.N., & Myaeng, S.H. (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing, 13*(4), 177-186.

Kontos, J., & Sidiropoulou, M. (1991). On the acquisition of causal knowledge from scientific texts with attribute grammars. *Expert Systems for Information Management, 4*(1), 31-48.

Message Understanding Conference. (1992). *Proceedings of the Fourth Message Understanding Conference (MUC-4).* San Mateo, CA: Morgan Kaufman.

Message Understanding Conference. (1994). *Proceedings of the Fifth Message Understanding Conference (MUC-5).* San Francisco: Morgan Kaufmann.

Message Understanding Conference. (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6).* San Francisco: Morgan Kaufmann.

Message Understanding Conference. (1998). *Proceedings of the Seventh Message Understanding Conference (MUC-7)* [Online]. Available: http://www.muc.saic.com/proceedings/muc_7_toc.html.

Mooney, R. J. (1990). Learning plan schemata from observation: Explanation-based learning for plan recognition. *Cognitive Science, 14*(4), 483-509.

Riloff, E., & Lehnert, W. (1994). Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems, 12*(3), 296-333.

Schank, R. C. (1982). *Dynamic memory*. New York: Cambridge University Press.

Schubert, L., & Hwang, C.H. (1989). An episodic knowledge representation for narrative texts. In R.J. Brachman, H.J. Levesque & Reiter, R. (Eds.), *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning* (pp. 444-458). San Mateo, Calif.: Morgan Kaufmann.

Selfridge, Mallory, Daniell, Jim, & Simmons, Dan. (1985). Learning causal models by understanding real-world natural language explanations. In *The Second Conference on Artificial Intelligence Applications: The Engineering of Knowledge-Based Systems* (pp. 378-383). Silver Spring, MD: IEEE Computer Society.

Soderland, S, Aronow, D., Fisher, D., Aseltine, J., & Lehnert, W. (1995). *Machine learning of text-analysis rules for clinical records* (Technical Report, TE-39). Amherst, MA: University of Massachusetts, Dept. of Computer Science.

Sowa, J.F. (1984). *Conceptual structures: Information processing in man and machine*. Reading, MA: Addison-Wesley.

Swanson, D.R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine, 30*(1), 7-18.

Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence, 91*, 183-203.

Swanson, D.R., & Smalheiser, N.R. (1998). Analysis of text linkages in MEDLINE titles as an aid to scientific discovery. In *AAAI Symposium on Link Analysis, Oct. 1998* [Online]. Available: http://kiwi.uchicago.edu/linkanalysis.html.

Theodorson, G. A., & Theodorson, A. G. (1979). *A modern dictionary of sociology*. New York: Barnes & Noble Books.

Christopher Khoo is an Assistant Professor in the Division of Information Studies at the Nanyang Technological University (NTU). He obtained his MSc in Library and Information Science from the University of Illinois at Urbana-Champaign and his PhD from Syracuse University. Prior to joining the faculty of NTU, he worked for 8 years at the National University of Singapore Library in various departments. His main research interest is in the application of artificial intelligence and natural language processing techniques to information retrieval. He can be contacted at: School of Applied Science, Nanyang Technological University, Blk N4 2A-32, Nanyang Avenue, Singapore 639798, or by email at: *assgkhoo@ntu.edu.sg*

Chan Syin is an Associate Professor in the discipline of Computer Engineering, School of Applied Science at the Nanyang Technological University. She obtained her Bachelor's Degree in Electrical Engineering from the National University of Singapore, and a PhD in Computer Science from the University of Kent, UK. Her research interests include image compression, object tracking in image sequence and multimedia information retrieval. She can be contacted at: *asschan@ntu.edu.sg*

Niu Yun is a graduate student in the School of Applied Science, Nanyang Technological University. She is doing research in the area of natural language processing and information extraction. She received her BEng from the Xi'an Jiao Tong University and her MASc in Information Science from the Documentation and Information Center of the Chinese Academy of Sciences, China. She can be contacted at: *p145521829@ntu.edu.sg*

Alyssa Ang is a systems analyst with Maxtor Peripherals Pte Ltd. She obtained her Bachelor's Degree in Information Technology from the University of Southern Queensland and her MSc (Information Studies) from the Nanyang Technological University. Her research interests are in Supply Chain Management and E-commerce. She can be contacted at: *alyssa_ang@maxtor.com*